



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING



SVALA:

Pseudonymizer service
for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING



SVALA:

SVA = SVenska som Andraspråk, Linking & Annotation

Pseudonymizer service for Swedish L2 essays

Elena Volodina, Samir Ali Mohammed, Arild Matsson



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING



SweLL –

research infrastructure for
Swedish as a Second Language



RIKSBANKENS
JUBILEUMSFOND

STIFTELSEN FÖR HUMANISTISK OCH
SAMHÄLLSVETENSKAPLIG FORSKNING



Swedish Learner Language

SweLL –

research infrastructure for
Swedish as a Second Language

SweLL promises (main)

1. Deliver a well-annotated (gold standard) corpus of L2 essays

- 600 essays, approx 100 per CEFR levels A1-C1 + 100 for control L1 learner corpus
- Incl manual correction annotation & manually checked linguistic annotation
- Make available for research (and public?)



```
graph = {
  "source": [
    {"id": "s0", "text": "I "},
    {"id": "s1", "text": "don't "},
    {"id": "s2", "text": "know "},
    {"id": "s3", "text": "his "},
    {"id": "s4", "text": "lives "},
    {"id": "s5", "text": ". " }
  ],
  "target": [
    {"id": "t0", "text": "I "},
    {"id": "t1", "text": "don't "},
    {"id": "t2", "text": "know "},
    {"id": "t3", "text": "where "},
    {"id": "t4", "text": "he "},
    {"id": "t5", "text": "lives "},
    {"id": "t6", "text": ". " }
  ],
  "edges": {
    "e-s0-t0": {"id": "e-s0-t0", "ids":
    "e-s1-t1": {"id": "e-s1-t1", "ids":
```

SweLL promises (main)

2. Set a platform (and workflow) for

- Continuous upload of new essays
- Manual correction annotation
- Automatic linguistic annotation

SweLL

Hem › datainsamling › Studenter › Lägg till student

Lägg till student

Swell_id:

Kön:

Födelseårsintervall:

Tid i Sverige:

Högsta examen:

FIRST LANGUAGES

First language: #1

Språk:

I don't know his lives .

I don't know where he lives .

I don't know his lives .

I don't know where:M he~his:F lives

<sentence id="8f7-8b5"> [Visa XML]

OO: Direkt objekt (akusativobjekt)

token	msd	lemma	lex	sense
Jag	PN. UTR. SIN. DEF. SUB	jag	jag..pn.1	jag..1
vet	VB. PRS. AKT	veta	veta..vb.1	veta..1
inte	AB	inte	inte..ab.1	inte..1
var	VB. PRT. AKT	vara	vara..vb.1	vara..1
han	PN. UTR. SIN. DEF. SUB	han	han..pn.1	han..1
bor	VB. PRS. AKT	bo	bo..vb.1	bo..1
.	MAD			

</sentence>

SweLL promises (main)

- Set a platform for browsing L2 essays
 - in concordance fashion (+parallel view)
 - In full text fashion

KWIC Statistik Ordbild

Antal träffar: 71 160

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ... » Gå till sida av 2847 Visa kontext

Äro UNDERBÄRTELSEN 2012

Omgiven av stress är det svårt för människor att hitta lycka.

er syn för människoögat – parningsleken har blivit en ren gruppvaldikt som måste vara synnerligen stressande för ådan – samlas gudingarna i flockar på tiotusental ute på öppna havet för att rugga.

Duktiga flickor blir sjuka av stress .

Associationssfären kring ordet " klitorata " och flygplatsmiljön, ger också t.ex. bilder av stress ,

Då fanns det ingen sådan stress ,

m att " allt som är trevligt är bra för magen ", men jag är säker på att magen mår bättre när man inte stressar

Det är lunchrusning i Londons metro och ungefär två miljoner hungriga Londonbor stress

– Det dyker alltid upp folk vid lunchdags, men det är inte samma stress som under läsaet.

Hör du till dem som inte gillar att stressa runt i butiker inför jul?

Då kunde vi ta det lite lugnare i andra halvlek och inte stressa i anfallen som i den första.

Bort från huvudstaden och stressen .

Ingen stressar i Hasse Ahlstrands bok och skiva för barn.

I stället för att ta stress över situationen där man fortfarande behövde en poäng för att ha sitt på det torra, fick KSF stressen .

– Vi borde våga plocka bort stressar med resor och fruktansvärd arbetstakt.

De som jobbar i affärslivet eller exportindustrin (och hämtar in pengar till Finland) stressande och man kan börja må dåligt.

som ung, då man ännu söker sig själv, möta alla krav och förväntningar kan kännas frustrerande och stress kan ge upphov till.

Clownerna konfronteras med det personliga mörker som stress .

Budskapet om fred och sinnesfrid går hand i hand med konfliktskylla känslor om materiella ting och stress .

Det fanns gott om tid så vi har inte behövt stress i onödan, konstaterar Kaikkonen som tillsammans med överstyrmannen ansvarar för att ta och konsumtionen i väst är ett verkligt alternativ, ett alternativ till ett meningsfullt arbete, a stressen

Filmen frågar om stress och trötthet.

Äro UNDERBÄRTELSEN 2013

– Det var stressande och jag var helt slut.

Eller med ökande brädska och stress ,

som ligger bakom, kanske det att jag mest åkt för mig själv och med fru Nives, och inte tagit någon stress ,

Osäkerhet, stress och knappa resurser är vardag för många kommunalt anställda.

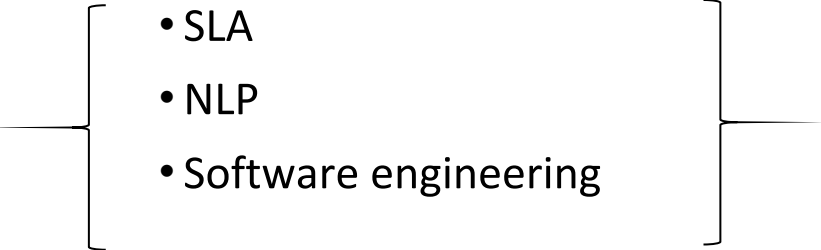
SW1203-UPPSATSER

I Sverige lever många människor det goda livet tycker jag. Det är inte så i många andra land. Människor bryr sig om att äta, träna och sova ordentligt. De vill också ha ett rikt socialt liv vilket är viktigt för psykosocial hälsa. Att ha ett gott liv är något viktigt för mig. Det är ett ständigt jobb. Man måste alltid tänka på sin hälsa. Om man inte har några problem med hälsa måste man träna. Idag sitter vi mycket mer än förrut. Sittande arbete gör oss lat och vi har inte inspiration att börja röra på oss. En familj som ofta vandrar i fjällen, cyklar, promenerar eller lekar ute tillsammans har det goda livet enligt mig. Mat är en av viktigaste saker angående det goda livet. Man måste välja väldigt noga sin mat. Det är lätt att vara nöjd med halvferdig mat vilket man lagga snabbt. Att lagga riktig nytig mat tar mycket tid och man bör förbereda sig. Jag menar att jag måste köpa färska grönsaker om jag ska lagga någon nytig mat. Jag menar med mat försöker man att äta hälsosamt och undvika fetma, diabetes, hög blodtryck och hjärt och kärlsjukdomar. Det är bäst om man är vegetarisk och icke-rökare. Jag tror att frasen "Det goda livet " ska referera till glad familjen som lever hälsosamt liv utan stress. Å andra sida är jag inte säkert att det är möjligt i ett modernt samhälle leva detta liv. I dagens samhället ar viktigt att tjäna mycket pengar därför att pengarna betyder en hög status och vi alla vill ha hög status.

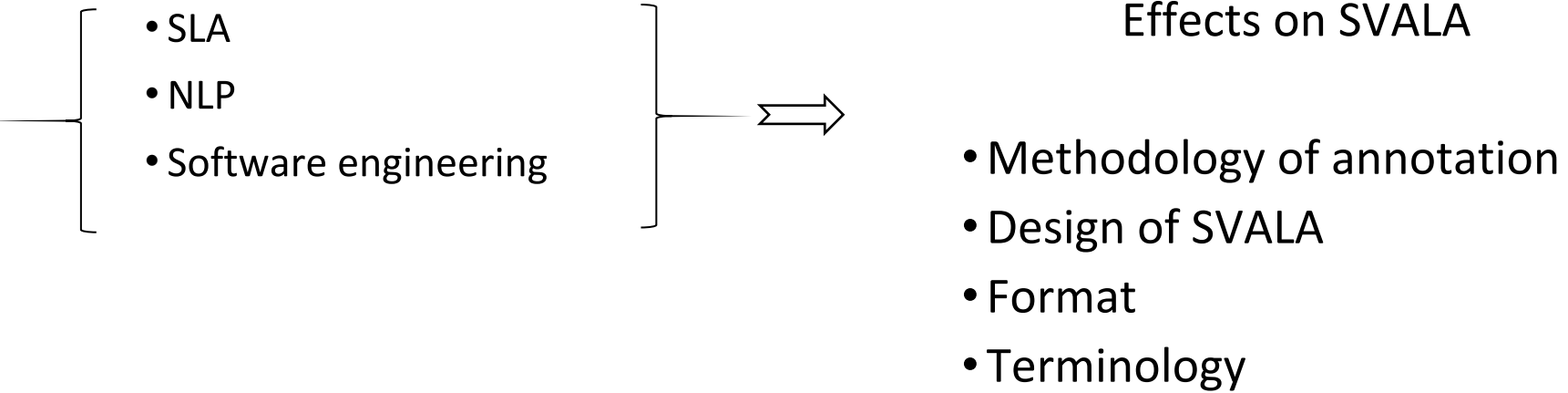
I modernt samhälle kommer tyvärr stress och många andra negativa saker. Till slutet vill jag säga att det goda livet är mitt mål. Ett foto av lycklig familjen på ett bord.

Nu tillbaka till Europa och Sverige. Här har människorna andra problem. Stress, långa

A truly interdisciplinary effort

- 
- SLA
 - NLP
 - Software engineering

A truly interdisciplinary effort



Methodology

- Anonymization in 3 steps:
 - detection
 - placeholder labelling
 - pseudonymization
- Normalization separate from correction annotation

undo redo Svala pseudonymization

previous next
prev mod next mod
Filter / numeric label

2 Morphology
gen
def

Alice and Bob went to Paris . Alice's wallet was stolen .
firstname_female 1 firstname_male 2 city 3 firstname_female 1 gen
Anna and David went to A-stad . Anna-gen wallet was stolen .

The screenshot shows a software interface for text pseudonymization. At the top, there are 'undo' and 'redo' buttons, followed by the title 'Svala pseudonymization'. Below this is a control panel with 'previous' and 'next' buttons, 'prev mod' and 'next mod' buttons, and a text input field labeled 'Filter / numeric label'. On the left side, there is a sidebar with a blue header '2 Morphology' and two buttons, 'gen' and 'def'. The main area displays a text transformation. The top line is the original sentence: 'Alice and Bob went to Paris . Alice's wallet was stolen .'. Below this, a row of placeholder labels is shown: 'firstname_female 1' under 'Alice', 'firstname_male 2' under 'Bob', 'city 3' under 'Paris', and 'firstname_female 1 gen' under 'Alice's'. The bottom line is the transformed sentence: 'Anna and David went to A-stad . Anna-gen wallet was stolen .'. Vertical lines connect the words in the top sentence to the placeholders, and the placeholders to the words in the bottom sentence.

Design

- As user-friendly (for the target user group) as possible:
 - No xml tags
 - Work in a "familiar" environment (i.e. text editor)
 - Visual support
 - Support for word-order changes

Svala normalization

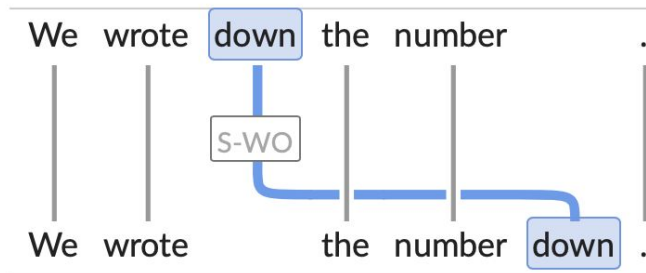
Source text:

We wrote **down** the number .

copy to target

Target text:

We wrote the number **down** .



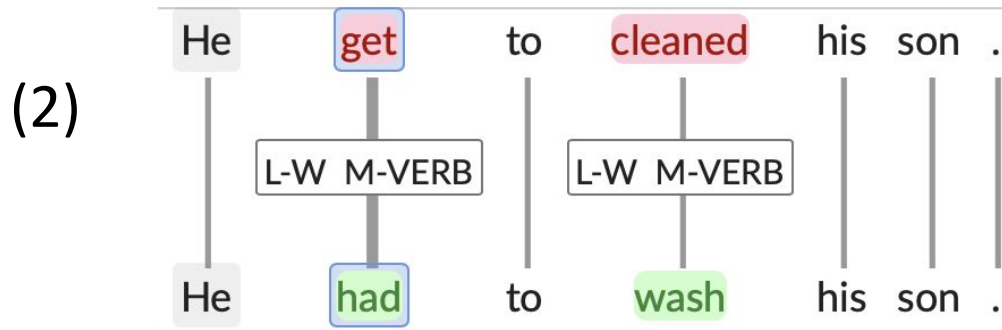
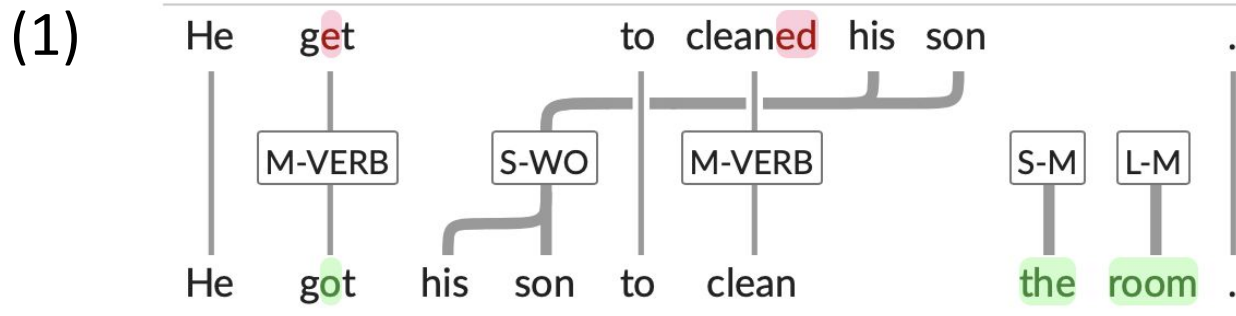
Format

- JSON format
- 3 separate "objects":
 - for original text
 - normalized text
 - links with labels

```
{
  "source": [
    {"id": "s0", "text": "We "},
    {"id": "s1", "text": "wrote "},
    {"id": "s2", "text": "down "},
    {"id": "s3", "text": "the "},
    {"id": "s4", "text": "number "},
    {"id": "s5", "text": ". "}
  ],
  "target": [
    {"id": "t0", "text": "We "},
    {"id": "t1", "text": "wrote "},
    {"id": "t2", "text": "the "},
    {"id": "t3", "text": "number "},
    {"id": "t4", "text": "down "},
    {"id": "t5", "text": ". "}
  ],
  "edges": {
    "e-s2-t4": {
      "id": "e-s2-t4",
      "ids": ["s2", "t4"],
      "labels": ["S-WO"],
      "manual": true
    },
  },
}
```

Terminology

- Correction annotation (not error annotation!)



SVALA - SweLL annotation tool

- Parallel text
- Visualized diff
- Semi-automatic word alignment
- Annotation on source–target links

Desired:

- Drop-down menus for error codes
- Three-tier representation (original, spell-corrected, normalized)
- Support with automatic spelling error detection



Dan Rosén,
research engineer



Arild Matsson,
research engineer

SVALA - automatic pseudonymizer service

- Why?
 - GDPR
 - Ethical reasons
 - To speed up annotation work
 - To boost essay collection

SVALA - automatic pseudonymizer service

- 8 head categories, 40 subcategories, morph. markers
 - names, geo names, institutions, transportation, age, date, miscellaneous, mark/sensitive

1. ORIGINAL TEXT → @PLACEHOLDER → RENDERING

2. ORIGINAL TEXT → @PLACEHOLDER → REPLACEMENT

3. ORIGINAL TEXT → @PLACEHOLDER → ORIGINAL
--

Guidelines: https://spraakbanken.github.io/swell-project/Anonymization_guidelines



Samir Ali Mohammed,
Systems developer



Arild Matsson,
research engineer

SVALA - pseudonymizer: facts

- Rule-based
 - tokenization, regular expressions
 - minor spelling correction (Levenstein distance)
- Git Repo: https://github.com/SamirYousuf/LR_project
- Python
- IN: text; OUT: json / text
- Three steps:
 - detection
 - placeholder labelling
 - pseudonymization

Yousuf Ali Mohammed Updated dataset	
..	
	Prof_dataset.csv
	_city_country_population.csv
	cities_sweden.csv
	city_country.csv
	city_country_population.csv
	country_capital.csv
	dict_data.json
	island_sweden.csv
	language.csv
	names_database.json
	names_database_1.json
	swedish_streets.csv

SVALA – hands-on demo

<https://spraakbanken.gu.se/swell/dev/>

Example essay (translation into English + mocking errors)

- I live in Stockholm on appartement . Jag är 29 . I live with my boyfriend . His name is Cezary . Apartement mine has a pattio and tree room . I enjoy there in Stockhulm but a lot of time to goto shop , fortifive minut . I have the buss and the Stockholm train . I lived in Danmark bifore , in Odense . It was less than Stockholm . I enjoy their too becaus I had more friends . I think it is hard to have friends here . But I enjoy better job here . In Odense jobbe I only on one website . In Stockholm I work on many website . I am webdevelooper . But Stockholm is closser to Luxembourg than Odense . It is important how one lives because I am not in my country . I mess my mother and my father but I live her with my boyfriend .

Future work

- Test on assistants → learn to incorporate automatically their corrections
- Integrate into Lärka and test on learners → collect their corrections into “reports”
- Adapt to English?
- Shared task?

Future work (for SVALA)

- Add (automatic) linguistic annotation and visualization
- Develop support for automatic error detection and labeling
- Make it available/modifiable for other languages/users (tagsets, versioning, server...)

GitHub repository

<https://github.com/spraakbanken/swell-editor>

A link to an article on SVALA

Wirén Mats, Arild Matsson, Dan Rosén, Elena Volodina. 2019. SVALA: Annotation of Second-Language Learner Text Based on Mostly Automatic Alignment of Parallel Corpora. *CLARIN-2018 post-conference volume*. LiUP Press.

<http://www.ep.liu.se/ecp/159/023/ecp18159023.pdf>

A link to an article on Anonymization

Beáta Megyesi et al. (2018). Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. *Proceedings of the 7th NLP4CALL workshop*. [[pdf](#)]

<http://www.ep.liu.se/ecp/152/006/ecp18152006.pdf>

A link to this presentation

<http://tiny.cc/gximdzt>

Thank you!

- Questions?
- Comments?

previous next
prev mod next mod

Filter / numeric label

1

Morphology

gen
def
pl

Names

firstname_male
firstname_female
firstname_unknown
surname
middlename
initials

Geographic data

country_of_origin
country
zip_code
region
city_swe
city
area
place
geo
street_nr

Institutions

institution
school
work
other_institution

Transportation

transport
transport_line

Age

age_digits

I live in Guntorp on apartment . I live with my boyfriend . His name

city_swe 1
I live in Guntorp on apartment . I live with my boyfriend . His name

is Hans . The apartment mine has a pattia and tree room . I

firstname_male 2
is Hans . The apartment mine has a pattia and tree room . I

enjoy there in Guntorp but a lot of time to goto shop , fortifive minut

enjoy there in Guntorp but a lot of time to goto shop , fortifive minut

. I have the bus and the Guntop train . I lived in Norway

transport 3 city_swe 1 transport 4 country 5
. I have the bus and the Guntorp train . I lived in Norway

bifore , in Tromsö . It was less than Gunntorp . I enjoy their too

Document comment:

- 1
- Guntorp
 - Guntop
 - Gunntorp
 - Guntorp
 - Guntorp

- 2
- Hans

- 3
- bus

- 4
- train

- 5
- Norway

- 6
- Tromsö

I live in **B-svensk-stad** in an apartment . I live with my boyfriend . His name is **Peter** . My apartment has a patio and three rooms . I enjoy it there in **B-svensk-stad** but it takes a lot of time to go to the shop , forty-five minutes . I have the **D-transport** and the **B-svensk-stad C-transport** . I lived in **C-land** before , in **D-stad** . It was smaller than **B-svensk-stad** . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In **D-stad** I worked only on one website . In **B-svensk-stad** I work on many websites . I am a web developer . But **B-svensk-stad** is closer to **B-land** than **D-stad** . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .

The apartment mine has a patio and tree room

My apartment has a patio and three rooms

undo redo

previous next

prev mod next mod

group orphan

auto

Enter filter text

L-M

L-R

L-REF

L-W

Morphological

M-ADJ/ADV

M-CASE

M-DEF

M-F

M-GEND

M-NUM

M-Other

M-VERB

Punctuation

P-M

P-R

P-W

Sent-Segmentation

Syntactical

S-adv

S-CON

S-finV

S-M

S-Msubj

S-R

S-W

S-WO

Unidentified

OBS!

Uni

SVALA correction annotation

show options

Document comment:

M-NUM

- room—rooms

O

- apartement—apartment
- pattio—patio
- tree—three

S-R

- The—

S-WO

- mine—My

Target text:

I live in B-svensk-stad in an apartment . I live with my boyfriend . His name is Peter . My apartment has a patio and three rooms . I enjoy it there in B-svensk-stad but it takes a lot of time to go to the shop , forty-five minutes . I have the D-transport and the B-svensk-stad C-transport . I lived in C-land before , in D-stad . It was smaller than B-svensk-stad . I enjoyed it there too because I had more friends . I think it is hard to find friends here . But I enjoy a better job here . In D-stad I worked only on one website . In B-svensk-stad I work on many websites . I am a web developer . But B-svensk-stad is closer to B-land than D-stad . It is important how one lives because I am not in my country . I miss my mother and my father but I live here with my boyfriend .

The apartment mine has a pattio and tree room .

My apartment has a patio and three rooms .

```

{
  "source": [
    {"id": "s18", "text": "The "},
    {"id": "s19", "text": "apartement "},
    {"id": "s20", "text": "mine "},
    {"id": "s21", "text": "has "},
    {"id": "s22", "text": "a "},
    {"id": "s23", "text": "pattio "},
    {"id": "s24", "text": "and "},
    {"id": "s25", "text": "tree "},
    {"id": "s26", "text": "room "},
    {"id": "s27", "text": "."}
  ]
}

```

SVALA [repo](#) [issues](#)

SVALA pseudonym. mode

[Demo](#)



undo redo

previous next

prev mod next mod

Filter / numeric label

2

Morphology

gen

def

Errors

ort

Names

firstname:male

firstname:female

firstname:unknown

surname

middlename

initials

Geographic data

country_of_origin

country

zip_code

region

city-SWE

city

area

place

geo

street_nr

Jag heter Ali och bor i Borlänge. Jag är 18 år. Jag

firstname:male 1

city-SWE 2

age_digits 3

Jag heter Peter och bor i Guntorp Jag är 19 år. Jag

flyttade till Sverige för 3 år sedan. Jag gillar Borlänges gator.

year 4

city-SWE 2 gen

flyttade till Sverige för 2 år sedan. Jag gillar Guntorp-gen gator.

show options

1

- Ali

2

- Borlänge.
- Borlänges

3

- 18

4

- 3

1. ORIGINAL TEXT → @PLACEHOLDER → RENDERING

2. ORIGINAL TEXT → @PLACEHOLDER → REPLACEMENT

3. ORIGINAL TEXT → @PLACEHOLDER → ORIGINAL

```
e-s6-t49": {
  "id": "e-s6-t49",
  "ids": ["s6", "t49"],
  "labels": ["city-SWE", "2"],
  "manual": true
}
```